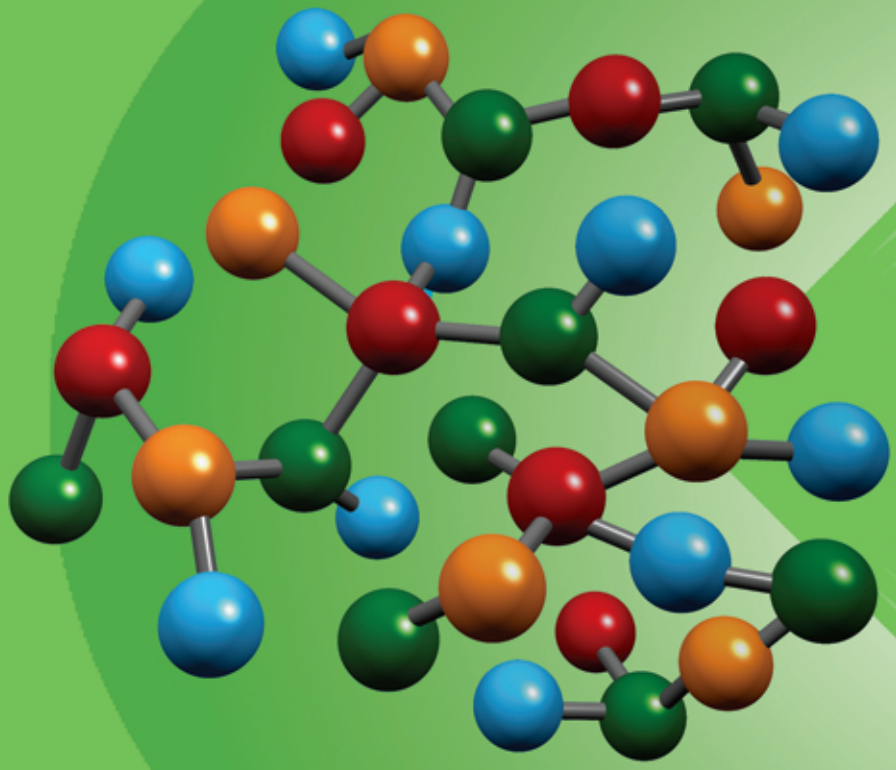
CEPOS INSILICO
SCIENCE FOR SCIENTISTS

10

USER MANUAL



PARAFIT™

© 2010. COPYRIGHT CEPOS INSILICO LTD

Impressum

Copyright © 2009 by CEPOS InSilico Ltd.

The Old Vicarage
132 Bedford Road
Kempston
BEDFORD, MK42 8BQ
www.ceposinsilico.com

Manual David Ritchie

Software David Ritchie

Layout **www.eh-bitartist.de**

TABLE OF CONTENTS

1 INTRODUCTION 4

2 BACKGROUND 5
2.1 Spherical Harmonic Superpositions 5
2.2 Rotational Correlations 6
2.3 Transformed SDF Files 7

3 PROGRAM USAGE 9
3.1 Fitting Mode 10
3.2 Multi-Molecule SDF Files 12
3.3 Scoring Functions and Property Options 13
3.4 Matrix Mode 13
3.5 Canonical Mode 15
3.6 Move Mode 17
3.7 Summary of Input and Output Files 17
3.8 Summary of Command Line Options 18

4 SUPPORT 22

5 REFERENCES 23

© CEPOS InSilico 2009

1 INTRODUCTION

ParaFit™ superposes and compares molecules using the spherical harmonic (SH) expansions of the molecular surface and local surface properties calculated by ParaSurf [1]. By exploiting the special rotational properties of the spherical harmonic basis functions [2], computation times can be reduced by several orders of magnitude compared to conventional shape matching algorithms [3, 4]. Hence the ParaFit™ module is an essential component of the ParaSurf suite for virtual high throughput screening studies where very large numbers of compounds need to be assessed.

ParaFit™ provides three main calculation modes. In the default “fitting” mode, ParaFit™ superposes one or more “moving” molecules onto a single “fixed” reference molecule. The program can also perform *all-versus-all* superpositions in which each molecule is superposed in turn onto all others. In this “matrix” mode, a table of distance scores is written out in a format suitable for subsequent clustering analysis, for example. In addition to superposing molecules, ParaFit™ may also be used to align molecules to the coordinate axes in order to place then in a standard or “canonical” orientation. This is often a useful first step in QSAR studies.

ParaFit™ can also apply arbitrary coordinate transformations to a given list of ParaSurf™, VAMP [5], or Mopac [6] SDF files. These transformations could be supplied as part of a processing pipeline by other superposition programs, *e.g.* ParaMatch™, that do not have the capability to rotate complex quantum mechanical (QM) properties such as quadrupole and octupole moments and atomic orbital charge density matrix elements. ParaFit™'s ability to rotate all of the orientation-dependent QM information in an SDF file eliminates the need to recalculate expensive QM quantities for new molecular orientations.

.

© CEPOS InSilico 2009

2 BACKGROUND

2.1 Spherical Harmonic Superpositions

SH molecular surface shapes are represented as radial expansions of the form:

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm}\, y_{lm}(\theta,\phi),$$

where $y_{lm}(\theta,\phi)$ are normalized real spherical harmonic functions and a_{lm} are the expansion coefficients. The parameters (θ,ϕ) are the usual spherical coordinates with respect to the centre of the harmonic expansion (CoH). ParaSurf™ normally sets the CoH to be equal to the molecular center of gravity (CoG). Because the spherical harmonics functions form a complete orthonormal spherical basis set, it can be shown that they transform amongst themselves under rotation according to [2, 7]:

$$y_{lm}(\theta',\phi') = \sum_{m'=-l}^{l} R_{m'm}(\alpha,\beta,\gamma)\, y_{lm'}(\theta,\phi),$$

where $R_{m'm}(\alpha,\beta,\gamma)$ are real Wigner rotation matrix elements expressed in terms of the Euler *z-y-z* rotation angles, (α,β,γ). Using this rotational property, it is straight-forward to show that a rotated SH expansion may be constructed from an unrotated expansion by rotating the original expansion coefficients [3]:

$$a'_{lm} = \sum_{m'=-l}^{l} R_{mm'}(\alpha,\beta,\gamma)\, a_{lm'}.$$

In order to calculate a superposition between a pair of molecules, ParaFit™ translates the CoH of the moving molecule (B) to that of the fixed reference molecule (A) and then searches for the rotation that minimizes the “distance” between the corresponding pairs of spherical harmonic expansions:

$$D_{EUCLIDEAN} = \int [(r_A(\theta,\phi) - R(\alpha,\beta,\gamma)\, r_B(\theta,\phi)]^2\, d\Omega.$$

By exploiting the orthonormality of the basis functions, this expression reduces to:

$$D_{EUCLIDEAN} = \left|\underline{a}\right|^2 + \left|\underline{b}\right|^2 - 2\,\underline{a}.\underline{b}',$$

© CEPOS InSilico 2009

where $\underline{b}'$ represents the vector of rotated SH expansion coefficients of the moving molecule, etc. We call this a Euclidean distance function due to its analogy to Euclidean distances in ordinary 3D space. This function has units of Å^2 and clearly depends on the relative size of the molecules being compared. However, when comparing multiple molecules, it is often convenient to use normalized distance or similarity functions in which identical molecules give a score of zero or unity, respectively. For example, dividing by the sum of the magnitudes of the SH shape vectors gives the Hodgkin similarity score:

$$S_{HODGKIN} = \frac{2\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2} = 1 - \frac{D_{EUCLIDEAN}}{|\underline{a}|^2 + |\underline{b}|^2}.$$

Similarly, ParaFit™ implements the Carbo and Tanimoto similarity functions as:

$$S_{CARBO} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|.|\underline{b}|},$$

$$S_{TANIMOTO} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}'}.$$

It is generally not obvious which of the above scoring functions is to be preferred. In our experience they all give good pairwise superpositions with L≥6 SH expansions. ParaFit™ uses the above Tanimoto function as its default similarity function.

For each of the above similarity functions, ParaFit™ allows a composite score to be calculated for an arbitrary combination of the SH surface shape and the four key ParaSurf™ local surface properties [**8**], namely molecular electrostatic potential (MEP), ionization energy (IE_L), electron affinity (EA_L), and polarizibility (α_L):

$$S = w_{SURFACE}S^{SURFACE} + w_{MEP}S^{MEP} + w_{IEL}S^{IEL} + w_{EA}S^{EA} + w_{\alpha_L}S^{\alpha_L},$$

where $w_{SURFACE}$ represents a user-defined weight factor, *etc.* In ParaFit™, all similarity score weight factors are normalized to unity before use. However, the Euclidean distance function may be selected if un-normalised or explicitly scaled property combinations are required.

2.2 Rotational Correlations

ParaSurf™ superposes molecules using a brute-force rotational search over the three Euler rotation angles. Conceptually, each moving molecule is rotated with respect to the fixed reference molecule, and the Euler rotation that gives the greatest similarity (or smallest distance) score is recorded. This is essentially a Fourier correlation search in Euler angle coordinates. However, because good

© CEPOS InSilico 2009

superpositions may be achieved using only low order harmonic expansions, it is not necessary to use fast Fourier transform (FFT) techniques to accelerate the calculation. Indeed, in our experience, the FFT is only of benefit when L>=16, which is considerably higher than the recommended default ParaFit™ value of L=6.

In addition to using low order correlation searches, ParaFit™'s superposition calculations are accelerated in two further ways. The first technique exploits the fact harmonic expansions to order L can have no more than L^2 local maxima. Hence, ParaFit™ initially uses relatively large angular search steps of around 8^o to cover the search space. In order to sample angular space evenly and efficiently, these angular samples are generated from the vertices of an icosahedral tessellation of the sphere. For a given angular step size, this gives around 30% fewer sample points than a naïve equi-angular grid [3]. Once the approximate location of maximum similarity has been identified, it is then refined using a localized grid search in steps of 2^o. Both angular step sizes may be adjusted by the user.

The second acceleration technique is used when comparing multiple molecules. Rather than separately rotating each of the moving molecules in turn, it is more efficient to rotate the SH expansions of only the reference molecule and to compare these against each of the moving molecules. Thus, relatively expensive SH rotations are applied to just one rather than N molecules. Once the optimal rotations have been found, the moving molecules are rotated using the inverse of the corresponding reference rotations. Using these techniques, a pair of molecules may be superposed in around 1/20s on a 1.8GHz Pentium Xeon processor, and computation times may be further reduced by a factor of up to 5 if multiple molecules are compared in a single ParaFit™ run.

2.3 Transformed SDF Files

Once the rotational orientation has been determined, ParaFit™ writes each of the moving molecules to a new SDF file. Multi-molecule SDF files are fully supported. Each new file contains rotated and translated instances of the original atom coordinates, point charges, molecular and atomic multipoles, charge density matrix elements and spherical harmonic expansion coefficients. ParaFit™ treats files with an extension of `.asd` as ParaSurf™ ASD files (*i.e.* anonymous SDFs). The main ParaSurf™ data blocks that are transformed by ParaFit™ are listed in **Table 1**. All other quantities in the new SDF file are copied without change from the original data.

Table 1: The main data blocks that ParaFit™ transforms after a superposition calculation.

SDF Data Block	Version
<> The atom coordinate table	The atom coordinate and bond tables in MDL format.
<VAMPBASICS>	The VAMP molecular heat of formation, HOMO and LUMO energies, and dipole moments.
<MOPACBASICS>	The MOPAC molecular heat of formation, HOMO and LUMO energies, and dipole moments.
<NAO-PC>	The natural atomic orbital point charges calculated by VAMP.
<DIPOL>	The molecular dipole moment, calculated by VAMP or Mopac.

© CEPOS InSilico 2009

<QUADPOL>	The molecular quadrupole moment, calculated by VAMP or Mopac.
<OCTUPOL>	The molecular octupole moment, calculated by VAMP or Mopac.
<ATOMIC MULTIPOLES>	The VAMP/Mopac atomic MEP multipoles.
<DENSITY MATRIX ELEMENTS>	The VAMP/Mopac atomic orbital density matrix elements.
<MOLECULAR_CENTERS>	The ParaSurf™ CoH and CoG coordinates.
<SPHERICAL_HARMONIC_SURFACE>	The ParaSurf™ SH surface shape coefficients.
<SPHERICAL_HARMONIC_MEP>	The ParaSurf™ SH MEP expansion coefficients.
<SPHERICAL_HARMONIC_EA>	The ParaSurf™ SH EA_L expansion coefficients.
<SPHERICAL_HARMONIC_IEL>	The ParaSurf™ SH IE_L expansion coefficients.
<SPHERICAL_HARMONIC_ALPHA(L)>	The ParaSurf™ SH α_L expansion coefficients.

© CEPOS InSilico 2009

3 PROGRAM USAGE

ParaFit™ is a command-line driven program, and is normally launched from a Unix terminal window or a Microsoft Windows command console. The program accepts a number of optional parameters to control the type of calculation, followed by the names of one or more SDF files. The basic command syntax is:

```
parafit [options] <sdf-files>
```

where square brackets represent optional parameters, and angled brackets represent required file names, respectively. The optional parameters generally have sensible defaults, and may often be abbreviated. The full list of ParaSurf™ command options is described in **Section 3.7**. By default, ParaFit™ creates new SDF files based on the names of the supplied input files, and it writes a record of each calculation to a log file (`parafit.log`). ParaFit™ does not write to the terminal unless errors are encountered. However, these behaviours may be changed using Unix-style option keywords. For example,

```
parafit -nolog -stdout <sdf-files>
```

suppresses the log file and directs all output messages to the “standard output” Unix terminal or Windows console. Similarly, the `-nostdout` option suppresses all standard output.

The main ParaFit™ operating modes are described in the following sections. These sections use some example SDF files which contain SH data blocks calculated by ParaSurf™ for four dopamine antagonists. Although ParaSurf™ SDF files typically follow certain naming conventions, the highly abbreviated file names listed in **Table 2** are used here for clarity. ParaFit™ automatically handles any files that use the GZIP (`.gz`) or BZIP2 (`.bz2`) file compression formats. There is essentially no limit on the number or size of SDF files, or the number of molecules in a multi-molecule SDF file, that can be processed. Multiple multi-molecule SDF files may be processed in a single ParaFit™ run.

Table 2: The four dopamine receptor antagonist examples used in this Manual.

Name	WDI Number	SDF File
Lorazapam	WDI-0030	`a.sdf`
Diazepam	WDI-0032	`b.sdf`
Temazepam	WDI-1451	`c.sdf`
Olanzapine	WDI-1416	`d.sdf`

© CEPOS InSilico 2009

3.1 Fitting Mode

In fitting mode, the first supplied SDF file is treated as the query or “fixed” molecule, and all subsequent molecules are treated as moving or “database” molecules which are to be fitted to the fixed query. Hence at least two SDF files must be given in this mode:

```
parafit query.sdf db1.sdf [db2.sdf …]
```

Each database molecule is first translated to the CoH of the query structure, and its CoH and CoG are updated accordingly. The best superposition is then found using the two-stage brute-force Fourier rotational search as described above. In order to indicate that the new files represent the original molecules transformed into the frame of the query, the ParaFit™ output files are named as **db1_query.sdf**, **db2_query.sdf**, *etc.* If a calculation produces only one SDF output file, it may be named explicitly using the **-write** option. However, the above file name generating rule must be used when processing multiple files. **Figure 1** shows the contents of the log file after superposing diazepam (**b.sdf**) onto lorezapam (**a.sdf**) using the default calculation:

```
parafit a.sdf b.sdf
```

Figure 2 shows the corresponding molecular superposition in which the new SDF file (**b_a.sdf**) contains a rotated diazepam molecule translated into the lorazepam coordinate frame.

```
--------------------------------------------------------------------------------
Parafit 09.05 log starting at Sat Jun 14 17:21:40 2008 on host wigner.
--------------------------------------------------------------------------------
Copyright (c) 2006-09 Dave Ritchie, University of Aberdeen.
Marketed under licence by CEPOS InSilico Ltd.

/home/staff/dritchie/parafit/parafit-8.05//bin/parafit.i586 -fit a.sdf b.sdf

--------------------------------------------------------------------------------
Parafit mode = Fitting: superpose database molecules to query file

Spherical hamonic order    = 6
Similarity score function  = Tanimoto
Angular search step        = 8.00 -> 500x45 = 22500 samples
Angular refinement step    = 2.00 -> 8x8+8 = 72 samples

Property weights: SURFACE = 1.00 MEP = 0.00 IEL = 0.00 EAL = 0.00 ALPHA = 0.00
--------------------------------------------------------------------------------
Detected 1010 MB main memory.
Reading a.sdf
a.sdf(1)[LORAZEPAM]
Reading b.sdf
b.sdf(1)[DIAZEPAM]
Input data read in 0.025 seconds
Estimating 318 Kb of input text data (159 Kb/molecule)
--------------------------------------------------------------------------------
Scoring...

Query: a.sdf(1)[LORAZEPAM]
    0.98375364  b.sdf(1)[DIAZEPAM]
Scored 1x1 in 0.064 seconds (15.57/sec)
```

Figure 1: An example ParaFit™ log file, **parafit.log**.

© CEPOS InSilico 2009

```
--------------------------------------------------------------------------------
Writing score summary file: parafit.pft
Query: a.sdf(1)[LORAZEPAM]
Collecting 1 molecules from b.sdf
Collecting b.sdf(1)[DIAZEPAM]
Writing b_a.sdf

Output files written in 0.043 seconds

Parafit done in a total of 0.13 sec.
Maximum memory allocation: 343 Kb.

Parafit 08.05 log stopping at Sat Jun 14 17:21:40 2009 on host wigner.
```

Figure 1 continued

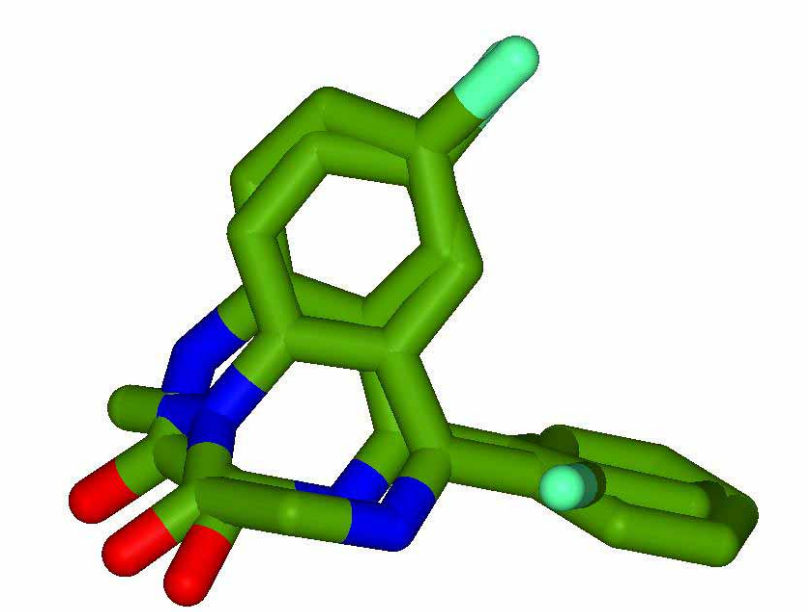


Figure 2: The superposition of lorazepam and diazepam, shown using the calculated diazepam orientation (b_a.sdf) in the coordinate frame of lorazepam (a.sdf).

The above example uses default values for the spherical harmonic expansion order (L=6), local surface property (surface shape), and similarity score function (Tanimoto). The same calculation could be specified more explicitly as:

```
parafit -fit -surface -tanimoto -order 6 -a1 8 -a2 2 <sdf's>
```

where **-a1** and **-a2** refer to the low and high resolution search increments, respectively.

In addition to the log file, ParaFit™ writes a compact summary of each run to a “scores” file, with a default file name of **parafit.pft**. This file lists the similarity scores and names of the original

© CEPOS InSilico 2009

query and database files. For example, superposing the remaining three dopamine antagonists onto lorezepam and explicitly naming the scores file using the **-score** option:

```
parafit a.sdf b.sdf c.sdf d.sdf -score abcd.pft
```

gives the scores file shown in **Figure 3**. The scores are always ordered by magnitude. Hence in this example, diazepam (**b.sdf**) is calculated to have the greatest surface shape similarity to lorazepam. If desired, the **-noscore** option may be specified to suppress the scores file.

```
1 3
Query: a.sdf(1)[LORAZEPAM]
     0.98375364  b.sdf(1)[DIAZEPAM]
     0.97739656  c.sdf(1)[TEMAZEPAM]
     0.96596095  d.sdf(1)[OLANZAPINE]
```

Figure 3: The contents of the example ParaFit™ scores file, abcd.pft, showing the calculated Tanimoto similarity scores for the three "database" molecules with respect to the given "query" molecule.

3.2 Multi-Molecule SDF Files

When dealing with more than just a handful of molecules, it is often convenient to collect multiple molecules in a single multi-molecule SDF file. In ParaFit™, all calculations are applied to each molecule in each SDF file. For example, the above calculation could equally have been performed as:

```
cat b.sdf c.sdf d.sdf >bcd.sdf

parafit -fit a.sdf bcd.sdf
```

In other words, **a.sdf** is treated as the "query" structure, which will be compared against the "database" of molecules in **bcd.sdf**. The result will be a new multi-structure SDF file called **bcd_a.sdf** which will contain each of the database molecules rotated into superposition with the query. These will be ordered by similarity with the query structure, with the most similar molecules appearing first. If desired, the original order may be maintained using the **-nosort** option.

When performing similarity searches against a large database, it is often desirable to retain only the most similar structures or "hits". In ParaFit™, this may be achieved using the **-hits** option. For example

```
parafit -fit -hits 20 query.sdf database.sdf
```

© CEPOS InSilico 2009

will return the 20 molecules (at most) in a file called **database_query.sdf** which most closely match the given query molecule. If desired, multiple "database" files may be searched in a single run. For example, the command

```
parafit -fit -hits 20 query.sdf db1.sdf db2.sdf
```

will produce two output files (**db1_query.sdf** and **db2_query.sdf**) which together contain the 20 molecules (at most) that most closely match the given query molecule.

If the first SDF file contains multiple molecules, then each of these molecules is treated as the query molecule in turn. Hence, for example, the commands

```
cat a.sdf b.sdf >ab.sdf

parafit -fit ab.sdf bcd.sdf
```

will produce two results files, **bcd_ab_1.sdf** and **bcd_ab_2.sdf**, which contain the three database molecules fitted to the first and second query molecules from **ab.sdf**, respectively.

3.3 Scoring Functions and Property Options

The Hodgkin, Carbo, or Tanimoto similarity scores, or Euclidean distance scores, may be selected using the options **-hodgkin**, **-carbo**, **-tanimoto,** or **-euclidean,** respectively. Similarly, the local property to use during the superposition may be specified using one of the options **-surface** (surface shape), **-mep** (MEP), **-iel** (IE_L), **-eal** (EA_L), or **-alpha** (α_L). Combinations of surface properties may be specified using the **-weights** option. For example:

```
parafit -carbo -weights -surface 0.8 -mep 0.2 <SDFs>
```

gives a Carbo superposition based on 80% shape similarity combined with a 20% contribution from the MEP. For the similarity scores, ParaFit™ normalises the weight factors to unity before use. Hence, the above calculation could be specified equivalently as:

```
parafit -carbo -weights -surface 80 -mep 20 <SDFs>
```

If un-normalised scores are required, the Euclidean scoring function should be specified:

```
parafit -euclidean -mep <SDFs>
```

3.4 Matrix Mode

In "matrix mode", each molecule is treated in turn as a query molecule, and all others are superposed onto it. For example, if N molecules are given as input, then N multi-molecule SDF files are produced as output, with each file containing the original query molecule plus the remaining N-1 molecules

© CEPOS InSilico 2009

superposed onto it. Thus, each output file correspond to one row of the NxN matrix . Hence, for example, the command

```
cat a.sdf b.sdf c.sdf d.sdf >abcd.sdf

parafit –matrix abcd.sdf
```

produces four output files named **abcd_matrix_1.sdf**, **abcd_matrix_2.sdf**, *etc.,* each of which will contain four molecules sorted by similarity with the query molecule. If desired, sorting can be suppressed by using the **–nosort** option.

In matrix mode, ParaFit™ writes an additional file of similarity scores in the "difference table" format of Kleiweg's publicly available clustering program [9]. For example,

```
parafit –matrix –hodgkin –dif abcd.dif abcd.sdf
```

generates a file of triangular distance scores (using D=1-S if necessary), as shown in **Figure 4**. **Figure 5** shows the dendrogram created from this file using the ParaShift™ utility script **dif2jpg**. This dendrogram readily confirms that olanzapine (**d.sdf**) is the outlier of the group. The option **–nodif** suppresses the difference file.

```
4
a.sdf(1)[LORAZEPAM]
b.sdf(1)[DIAZEPAM]
c.sdf(1)[TEMAZEPAM]
d.sdf(1)[OLANZAPINE
0.01624636          # ba = ab
0.02260344          # ca = ac
0.01412539          # cb = bc
0.03403905          # da = ad
0.03975538          # db = bd
0.02909502          # dc = cd
```

Figure 4: Example ParaFit™ output file of difference (distance) scores, abcd.dif. Any text following a hash (#) comment character is provided only to indicate the order in which the distance values appear in the file and is ignored by the clustering program.

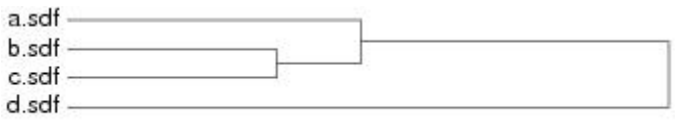


Figure 5: The dendrogram created from abcd.dif using the ParaShift™ dif2jpg utility.

If the main aim of a matrix mode calculation is to conduct a cluster analysis, it is often worthwhile using the **–nosdf** option to prevent ParaFit™ creating any SDF files. This avoids filling the working directory with a large number of unwanted files and helps to speed up the calculation. Depending on CPU speeds, ParaFit™ can perform rotational superpositions from 2 to 10 times faster than the time

© CEPOS InSilico 2009

required to write a new SDF file. Hence the speed-up can be considerable when clustering large datasets.

3.5 Canonical Mode

In canonical mode, ParaFit™ places each molecule in a standard or “canonical” orientation such that its maximal radial extent is aligned with the positive z axis, and (whilst keeping this axis fixed) its maximal equatorial extent is aligned with the positive x axis. For spherical harmonic expansions to order L=2, this corresponds to aligning molecules to the coordinate axes using their ellipsoidal radii or moments of inertia, for example. However, such low order alignments are ambiguous with respect to 180^{o} flips about the coordinate axes. Therefore, by default, ParaFit™ calculates canonical alignments using SH expansions to L=6 in order to eliminate any ambiguity in the final orientation (except for the rare cases of molecules with intrinsic C_{2V} symmetry). The syntax for canonical alignments is:

```
parfait -canonical moving1.sdf [moving2.sdf …]
```

As before, the SH surface shape function is used by default, although any local surface property may be used. In canonical mode, molecules are always aligned by maximizing the un-normalised SH property values with respect to the axes, regardless of any command line scoring or weighting options. The values subsequently written to the scores file are the magnitudes of the corresponding SH property vectors calculated at the current SH expansion order. Hence, by default, the scores file orders canonicalised molecules by surface area. **Figure 6** shows the four example dopamine antagonists in their L=6 canonical orientations.

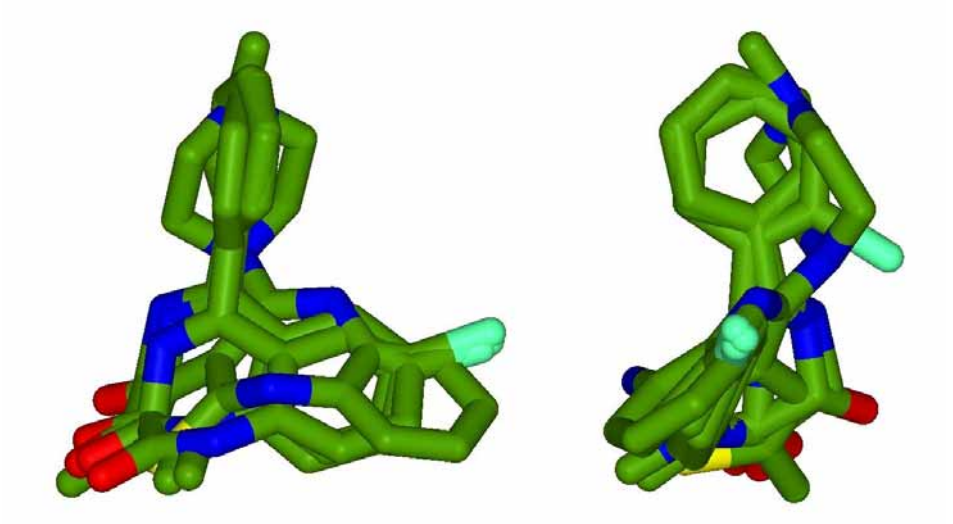


Figure 6: Left: the L=6 SH surface shape canonical orientations of lorazepam, diazepam, temazepam, and olanzapine; right: the same molecules rotated by 90o about the z axis.

Most operating systems limit the number of characters allowed in a command line, and this implicitly limits the maximum number of SDF files that may be specified using the command line syntax. In order to circumvent this limit, the `-read` option may be used to direct ParaFit™ to read the list of

© CEPOS InSilico 2009

SDF files to be processed from a specific file, with each file name in the list file being separated by space or newline characters. For example, the command:

```
parfait -canonical -read 74.lis
```

canonicalises the orientations of 74 selected drug molecules listed in the file **74.lis**. **Figure 7** shows the resulting orientations. This figure was produced with the help of two ParaShift™ Unix utility scripts, `sdf2pdb` and `pdb2one`, in order to concatenate the SDF files into an single PDB file for display using *Hex* [**10**].

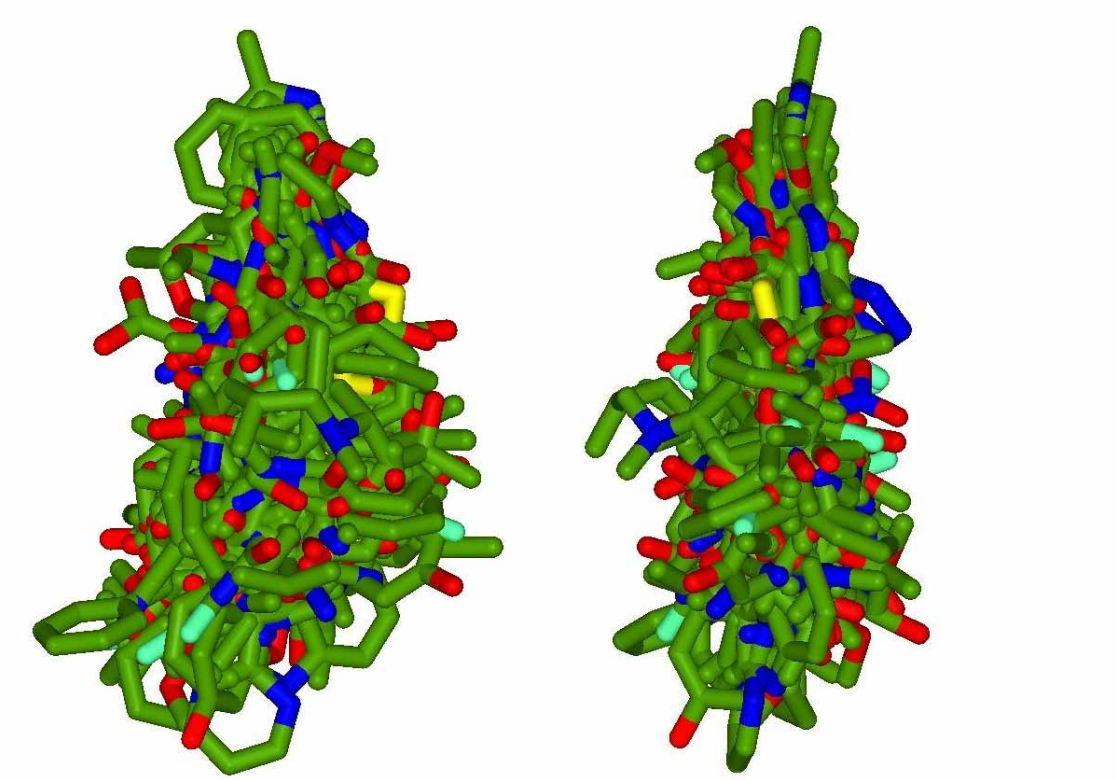


Figure 7: Left: 74 selected drug molecules in their L=6 canonical orientations; right: the same molecules rotated by 90o about the z axis.

Because no reference file is used in canonical mode, the new output file names are generated with a default affix of "canonical". Hence the above example would create files of the form: **moving1_canonical.sdf**, **moving2_canonical.sdf**, etc. The **-affix** option may be used to explicitly specify how the output files should be named. For example, the command:

```
parfait -canonical -read 74.lis -affix c
```

produces **moving1_c.sdf**, **moving2_c.sdf**, *etc.*

© CEPOS InSilico 2009

3.6 Move Mode

In “move mode”, no superpositions are calculated. Instead, each molecule is transformed according to a given sequence of rotation and translation operations. For example, the command:

```
parfait –move –ry 90 –tx 10 moving1.sdf [moving2.sdf …]
```

rotates each molecule relative its CoH by 90^{O} about the y axis, and then translates the result by 10Å in the *x* direction. ParaFit™ uses the convention that a positive rotation angle defines an anticlockwise rotation of the molecule as seen when looking along the axis of rotation towards the origin. All coordinate transformations (**`-rx, -ry, -rz, -tx, -ty, -tz`**) are applied in the order in which they are appear on the command line. It is also possible to specify that each molecule should have its CoH shifted to the global coordinate origin using the **`–move-coh`** option. For example,

```
parfait –move-coh –ry 90 –tx 10 moving1.sdf [moving2.sdf …]
```

rotates and co-locates the CoH of each molecule 10Å along the positive *x* axis. For completeness, ParaFit™ also allows coordinate operations to be applied relative to the CoG. For example, the command

```
parfait –move-cog moving1.sdf [moving2.sdf …]
```

moves all molecules such that their CoGs lie at the global coordinate origin. Conversely, specifying

```
parfait –move+cog –ry 90 moving1.sdf [moving2.sdf …]
```

rotates all molecules about their individual CoGs. From the above descriptions, the **`–move`** option is seen to be an abbreviation for **`–move+coh`** (*i.e.* move relative to CoH).

Like canonical mode, “move mode” adds an automatic file name affix to all output files, but in this case the default affix is “parafit”. As before, this may be changed by the **`-affix`** option. For example,

```
parfait –affix 0 –move-coh moving1.sdf [moving2.sdf …]
```

produces **`moving1_0.sdf`**, **`moving2_0.sdf`**, *etc.*

3.7 Summary of Input and Output Files

ParaFit™ assumes all input data files are SDF-format files. There is no requirement that SDF data files are named with the **`.sdf`** extension. However, anonymous SDF files (ASD files) must have an extension of `.asd` in order to be processed properly. ParaFit™ constructs SDF output file names from the given input file names. Hence if an input file name uses upper case, then so too will the corresponding output file. If an input file does not have an extension, then neither will the output file. However, if ParaFit™ fails to open a file with no extension, it will append **`.sdf`** to the name and try again. Any file names that end with **`.gz`** or **`.bz2`** are presumed to be compressed files; compressed

© CEPOS InSilico 2009

input files will cause compressed output files to be generated. In addition to the above rules, ParaFit™ writes up to three files of additional information, as described in **Table 3**.

Table 3: Summary of ParaFit™ additional information output files.

File Name	Description
`parafit.log`	Each ParaFit™ run creates a log file. Any existing file of the same name is overwritten. The log file may be named explicitly using the `-log` option. The log file may be suppressed using the `-nolog` option.
`parafit.pft`	Each ParaFit™ fit, matrix, or canonical mode calculation creates a similarity score file of this name. Any existing score file is overwritten. The score file may be named explicitly using the `-score` option. The score file is suppressed by `-noscore`.
`parafit.dif`	Each ParaFit™ matrix mode run creates a “difference” (distance) file of this name. Any existing difference file is overwritten. The file may be named explicitly using the `-dif` option. The difference file is suppressed by `-nodif`.
`parafit.csv`	Each ParaFit™ matrix mode run writes the calculated similarity (or distance) matrix to a “comma-seperated-values” (or “csv”) file of this name. Any existing csv file is over-written. The file may be named explicitly using the `-csv` option. The csv file is suppressed by `-nocsv`.

3.8 Summary of Command Line Options

Table 4 describes the ParaFit™ command line option keywords. A shorter version of these descriptions may be produced directly from the program using the `-help` option.

Table 4: List of ParaFit™ command line options.

Command Line Option	Description
`-fit` `-f`	Superpose one or more SDF files to a given reference SDF file. This is the default mode of calculation. The first SDF file is treated as the fixed reference molecule. All subsequent SDF files are treated as moving molecules to be fitted to the reference structure. At least two SDF files must be given in this mode. This option may be abbreviated as shown.
`-matrix` `-m`	Superpose two or more molecules in an *all-versus-all* manner to produce a “matrix” of superpositions. Each output SDF file name is constructed from corresponding pairs of moving and fixed reference file names. At least two SDF files must be given in this mode. This option may be abbreviated as shown.
`-canonical` `-c`	Align one or more molecules with the coordinate axes. Each output SDF file name is constructed by appending the source file name with the default affix of “canonical”. At least one SDF file must be given in this mode. This option may be abbreviated as

© CEPOS InSilico 2009

Command Line Option	Description
	shown.
`-order L` `-l L`	Set the spherical harmonic expansion order to use for superposition and canonicalization calculations. The default value is 6. This option may be abbreviated to `-l` (ell).
`-hodgkin`	Perform superpositions by maximizing the Hodgkin similarity score. This is the default score function.
`-carbo`	Perform superpositions by maximizing the Carbo similarity score.
`-tanimoto`	Perform superpositions by maximizing the Tanimoto similarity score.
`-euclidean`	Perform superpositions by minimizing the Euclidean distance score.
`-weights` `-w`	Superpose molecules using multiple local surface properties, using a given numerical weight factor for each property. If no property keywords are given, the calculation is performed as if `-surface 1.0` had been specified. This keyword may be abbreviated as shown.
`-noweights` `-now`	Do not superpose multiple SH properties with user-supplied weights. Instead, molecules will be superposed using a single property, with surface shape (`-surface`) as the default property. The default behavior is to superpose using a single property (*i.e.* as if `-noweights` had been specified). This keyword may be abbreviated as shown.
`-surface` `-surface W`	Superpose molecules using SH molecular surfaces (this is the default superposition property). If the `-weights` option has been specified, a numerical weight factor W must be provided after this option keyword. If the `-noweights` option has been specified, no weight factor should be given.
`-mep` `-mep W`	Superpose molecules using the SH molecular electrostatic potential (MEP), or include the MEP with the given weight factor W in a multi-property score, as described above.
`-iel` `-iel W`	Superpose molecules using the SH local ionization energy (IE_L), or include the IEL with the given weight factor W in a multi-property score, as described above.
`-eal` `-eal W`	Superpose molecules using the SH local electron affinity (EA_L), or include the EAL with the given weight factor W in a multi-property score, as described above.
`-alpha` `-alpha W`	Superpose molecules using the SH polarizability (α_L), or include the polarizability with the given weight in a multi-property score, as described above.

`-angle A`
`-a A`
`-a1 A`
	Set the angular step size A for the first pass rotational search. The default value is 8°. This option may be abbreviated as shown.

`-angle2 A` `-a2 A`	Set the angular step size A for the second pass rotational refinement search. The default is 2°. This option may be abbreviated as shown.
`-read file`	Read a list of SDF input file names from the given file.
`-write file`	Specify the explicit name of a single SDF output file. This is only permitted when there

© CEPOS InSilico 2009

Command Line Option	Description
	is just one output file.
-score file	Write similarity scores to a given file (the default file name is **parafit.pft**).
-noscore	Suppress the scores file.
-dif file	In matrix mode, write distance scores to a given “difference” format output file (the default file name is **parafit.dif**).
-nodif	Suppress the difference (distance) file.
-log file	Write all messages to a named log file (the default log file is **parafit.log**).
-nolog	Suppress the output of a log file.
-stdout	Write all messages to the Unix terminal or Microsoft Windows command console (“standard output”).
-nostdout	Suppress writing messages to standard output (this is the default).
-sdf	Write new SDF files after a superposition or canonicalization calculation. This is the default.
-nosdf	Suppress the output of new SDF files.
-affix F	When generating new SDF files in canonical or move mode, construct output file names by inserting the given affix F between the root and extension components of the input SDF file names. In canonical mode, the default affix is “canonical”. In move mode, the default affix is “parafit”.
-move-coh	Move one or more SDF files to locate their harmonic expansion centers (CoHs) at the origin, and apply any subsequent transformations (**-rx**, **-tx**, *etc.*) relative to this new origin in the order in which they appear on the command line.
-move-cog	Move one or more SDF files to locate their centers of gravity (CoGs) at the origin, and apply any subsequent command line transformations (**-rx**, **-tx**, *etc.*) relative to this new origin in the order in which they appear on the command line.
-move+coh **-move**	Apply a given sequence of transformations (**-rx**, **-tx**, *etc.*) in the order in which they appear on the command line to each molecule relative to the individual molecular CoHs. This option may be abbreviated as shown.
-move+cog	Apply a given sequence of transformations (**-rx**, **-tx**, *etc.*) in the order in which they appear on the command line to each molecule relative to the individual molecular CoGs. This option may be abbreviated as shown.
-rx X	Apply an anticlockwise rotation of X degrees about the x axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.
-ry Y	Apply an anticlockwise rotation of Y degrees about the y axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.

© CEPOS InSilico 2009

Command Line Option	Description
`-rz Z`	Apply an anticlockwise rotation of Z degrees about the z axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.
`-tx X`	Apply a translation of X Ångstroms along the x axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.
`-ty Y`	Apply a translation of Y Ångstroms along the y axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.
`-tz Z`	Apply a translation of Z Ångstroms along the z axis to the current orientation of each molecule in the given list of SDFs. The initial coordinate origin is selected using one of the above **move** options.
`-debug`	Produce verbose debugging output.
`-version` `-v`	Print the program version number. This option may be abbreviated as shown.
`-help`	Print a summary of all ParaFit™ program options.

© CEPOS InSilico 2009

4 SUPPORT

Any questions regarding ParaFit™ should be sent to **support@ceposinsilico.com**

© CEPOS InSilico 2009

5 REFERENCES

1 J.-H. Lin, T. Clark, *An Analytical, variable resolution, complete description of static molecules and their intermolecular binding properties*, J. Chem. Inf. Model, 2005, 45, 1010-1016.

2 M.E. Rose, *Elementary Theory of Angular Momentum*, 1957, Wiley, New York.

3 D.W. Ritchie, G.J.L. Kemp, Fast Computation, *Rotation, and Comparison of Low Resolution Spherical Harmonic Molecular Surfaces*, J. Comp. Chem. 1999, **20**(4), 383-395.

4 D.W. Ritchie, G.J.L. Kemp, *Protein Docking Using Spherical Polar Fourier Correlations*, Proteins: Struct. Funct. Genet. 2000, **39**, 178-194.

5 T. Clark, A. Alex, B. Beck, F. Burkhardt, J. Chandrasekhar, P. Gedeck, A.H.C. Horn, M. Hutter, B. Martin, G. Rauhut, W. Sauer, T. Schindler, and T. Steinke, VAMP 8.2, 2002; available from Accelrys Inc., San Diego, USA.

6 J.J.P. Stewart, MOPAC2000, 1999, Fujitsu Ltd., Tokyo, Japan. MOPAC 6.0 was once available as: J.J.P. Stewart, QCPE #455, Quantum Chemistry Program Exchange, Bloomsville, Indiana, 1990.

7 L.C. Biedenharn, J.C. Louck, *Angular Momentum in Quantum Physics*, 1981, Addison-Wesley, Reading, MA.

8 B. Ehresmann, M.J. deGroot, A. Alex, and T. Clark, *New Molecular Descriptors Based on Local Properties at the Molecular Surface and a Boiling-Point Model Derived from Them*, 2004, J. Chem. Inf. Comp. Sci. **44**, 658-668.

9 **http://www.let.rug.nl/~kleiweg/clustering/**;

10 **http://www.csd.abdn.ac.uk/hex/**

© CEPOS InSilico 2009